Linear_regression_preliminary_analysis

Team: Who's there

Thursday, August 18, 2016

Introduction

The aim of the project was to be able to find a model that best predict the relationship between: * the number of people counted with the survey for a given class at a particular hour * Wi-fi log counted in that room at that hour

This will allow to see whether Wi-Fi log is a good predictor for estimating occupancy in a classroom.

We first tried to see if the relationship between this two variables was linear. To do so we run a linear regression.

Below we describes step by step all the analysis performed.

ANALYSIS

DATABASE CONNECTION AND DATASET

```
## Loading required package: DBI
```

First of all, we set up the connection to the database, using the following code:

```
##connection <- dbConnect(MySQL(),user="root", password="",dbname="who_there_db", host="localhost")
```

Then we made a query to the database, in order to get all the groundth truth data collected in room B.002, B.004 and B.006 from 9 to 17 and the correspondent Wi-Fi Log measured in that time frame and rooms.

The dataset created had in total 216 rows and it will allow us to explore if Wi-Fi log is a good predictor of the observed occupancy of the room in a certain hour.

As a **target features** for our linear regression we decided to use the number of associated client, calculated multiplying the percentage of the room full with the capacity of the room.

As response variables or feature we considered Wi-Fi logs, which were summarised either as average of the logs counted for each room and for each hour or as maximum of the logs measured for each room and for each hour.

Together with the Wi-Fi log, we included in the data set the following features:

- **Date**, which we did not use in this analysis, because they just cover 2 weeks of Novemeber, but for future analyses they can be used to group observations by seasons or semesters or to finds seasonal trends for time series analyses.
- **Time**, which will be explored either as continous variable and as categorical to explore if the time of the day can have an affect on the Wi-Fi log. To do so we, bin the time in 4 ranges: early morning (9-11), late morning (11-13), early afternoon (13-15) and late afternoon (15-17). This will allow us to see if the Wi-Fi log accuracy was changing during the day. For example, it is more likely that all the elctronic devices are fully powered early in the morning and consequently the Wi-fi log data can be more accurate or overestimating the occupancy of the room (i.e. more than one device per person). On the contrary in the afternoon, the devices may be more likely to be out of battery and it is possible that there are less devices in the room.

- **Module**, which we are not going to include in the analysis because the majority of the module present are for computer science, but for future analyses it will be possible to explore if the Wi-fi log accuracy in predicting the occupancy change acroos the courses. Science course or computer science course will more likely to use electronic devices during lecture than art students.
- **Course level**, which can indicate us whether electronic devices will be less used during different course level. For example, first and second level courses can be less practical and therefore laptop are not needed and that can decrease the number of devices connected. On the other hand, undergraduate might be more distracted during lecture and look at their phones during lectures. This will result in an increase of connection in that hour.
- **Tutorial**, which can affect the number of logged people. First of all, because tutorial divided the room in 2 and therefore there will be measured less people than expected.
- **Double_module**, categorical variable indicating wether in the class there are more than one module, increasing the number of people expected in the room.
- **Double_module**, categorical variable indicating wether in the class went ahead to check for false positive.

The resulting data set is printed below:

```r
head(AnalysisTable)
```

```
##   Room       Date Time    Module Course_Level Tutorial Double_module
## 1    1 2015-11-03    9         0            0        0             0
## 2    1 2015-11-04    9 COMP30190            3        0             0
## 3    1 2015-11-05    9         0            0        0             0
## 4    1 2015-11-06    9 COMP30220            3        0             0
## 5    1 2015-11-09    9 COMP30190            3        0             0
## 6    1 2015-11-10    9         0            0        0             0
##   Class_went_ahead Capacity Percentage_room_full Wifi_Average_logs
## 1                1       90                 0.00            4.7500
## 2                1       90                 0.25           13.4545
## 3                1       90                 0.00            6.8333
## 4                1       90                 0.00            2.4167
## 5                1       90                 0.25           14.7273
## 6                1       90                 0.00            2.2727
##   Wifi_Max_logs Survey_occupancy   Factor_Time
## 1            21              0.0 Early Morning
## 2            15             22.5 Early Morning
## 3            29              0.0 Early Morning
## 4             3              0.0 Early Morning
## 5            18             22.5 Early Morning
## 6            14              0.0 Early Morning
```

DATA QUALITY REPORT

Before running any analyses, we carried out the data quality report to check for any issue related to the variable (e.g. outlier, skewed distribution, NaN values) and solutions we will implement to solve them.

Initially set all the categorical variables as factors and then we printed the descriptive statistic for all the features.

```
summary(AnalysisTable)
```

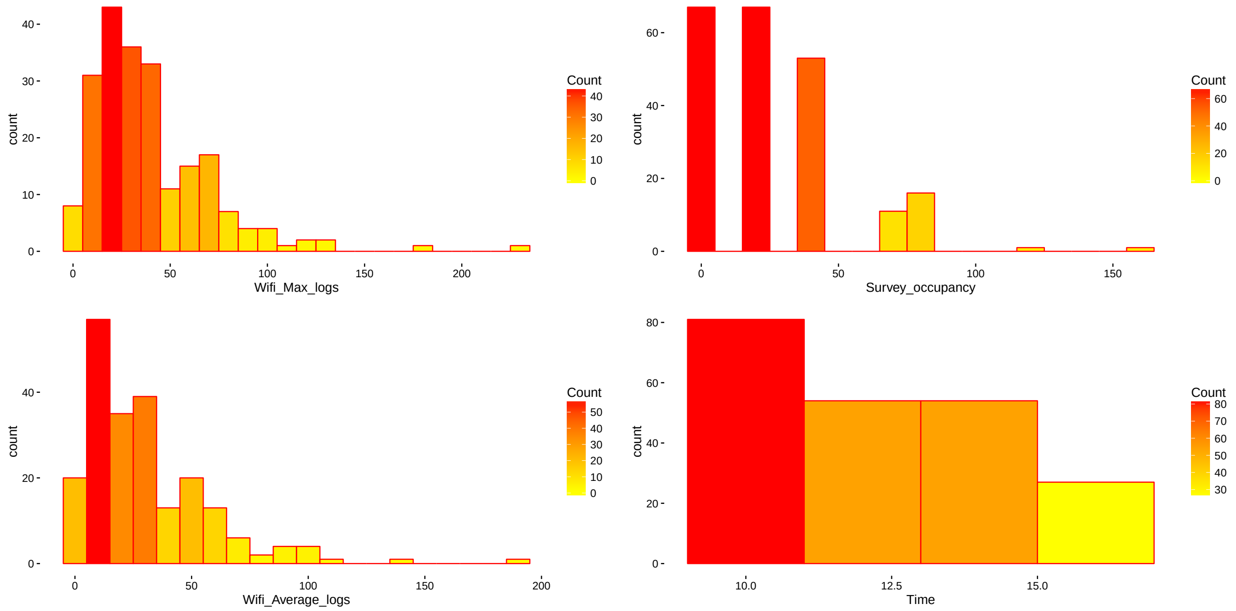
```
##       Room           Date         Time             Module    Course_Level
##  Min.   :1   2015-11-03:24   Min.   : 9.00   0        : 59   0:59
##  1st Qu.:1   2015-11-04:24   1st Qu.:10.75   COMP30080: 12   1:14
##  Median :2   2015-11-05:24   Median :12.50   COMP47300: 12   2:23
##  Mean   :2   2015-11-06:24   Mean   :12.50   COMP47290: 10   3:76
##  3rd Qu.:3   2015-11-09:24   3rd Qu.:14.25   COMP20020:  9   4:40
##  Max.   :3   2015-11-10:24   Max.   :16.00   COMP30240:  8   5: 4
##              (Other)   :72                   (Other)  :106
##     Tutorial       Double_module Class_went_ahead    Capacity
##  Min.   :0.00000   0:210         0: 22            Min.   : 90.0
##  1st Qu.:0.00000   1:  6         1:194            1st Qu.: 90.0
##  Median :0.00000                                  Median : 90.0
##  Mean   :0.02778                                  Mean   :113.3
##  3rd Qu.:0.00000                                  3rd Qu.:160.0
##  Max.   :1.00000                                  Max.   :160.0
##
##  Percentage_room_full Wifi_Average_logs Wifi_Max_logs    Survey_occupancy
##  Min.   :0.00         Min.   :  0.00    Min.   :  0.00   Min.   :  0.00
##  1st Qu.:0.00         1st Qu.: 11.61    1st Qu.: 18.75   1st Qu.:  0.00
##  Median :0.25         Median : 23.67    Median : 32.50   Median : 22.50
##  Mean   :0.25         Mean   : 30.33    Mean   : 40.01   Mean   : 28.01
##  3rd Qu.:0.25         3rd Qu.: 43.88    3rd Qu.: 55.25   3rd Qu.: 40.00
##  Max.   :1.00         Max.   :192.92    Max.   :230.00   Max.   :160.00
##
##           Factor_Time
##  Early Morning  :54
##  Late Morning   :54
##  Early Afternoon:54
##  Late Afternoon :54
##
##
##
```

From this we could see that NaN values were not present in the data set. We could notice that the observations for the features Tutorials and Double_model were not even distributed across the 2 levels of the variables. In fact, only 6 observations were present for tutorial class and for double module class. Therefore, we decided to discard both the features, because they will be not informative for the analysis. Similarly for the feature class_went_ahead the majority of the lectures did occur and we decided to discard it. Furthermore, for the variables Wifi_Average_clients, Wifi_Max_clients and Survey_counted_clients it seems that there are few outliers, since the median is lower than the mean and the max values are far higher than the mean values. We will going to explore this issues with histogram and boxplots.

Exploratory graphs

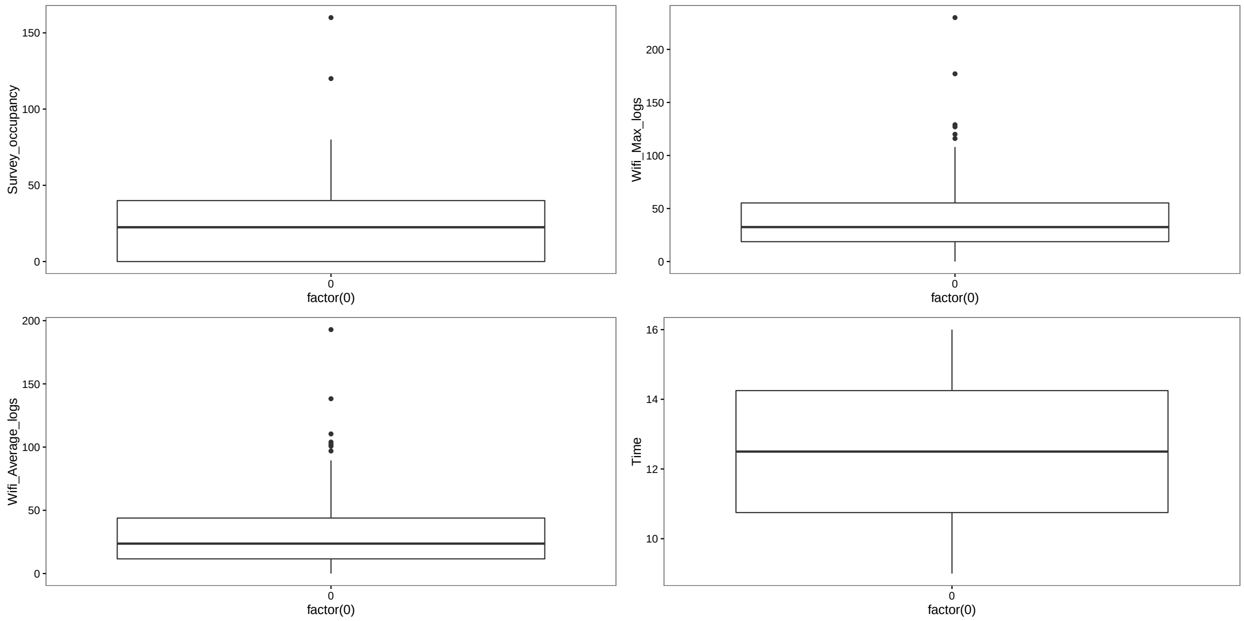
For exploring possible issues related with the continuous variables we plotted histograms and boxplots.

Histograms



Form the histograms we could see that the distribution of the feature Wifi Maximum_client (i.e. the Maximum number of devices logged in one hour lecture) was skewed to the left, indicating that the in the majority of the lecture were counted no more than 40 people. Furthermore, we could see that there are potential outliers (values $>$ 150). Similar pattern was observed for the feature Wifi_Average_clients. Different was the situation of the target feature, Survey_counted client, which showed a skewed distribution, but more scattered, similar to a Poisson distribution. This can cause a problem in running a linear regression and more likely we have have to run a generalise linear model with a Poisson distribution. This is not surprising, since we are dealing with count data (Zuur et al. 2009). Feature times had as well a skewed distribution, suggesting that the majority of the lectures were concentrating during the early morning and they were decreasing towards the afternoon.

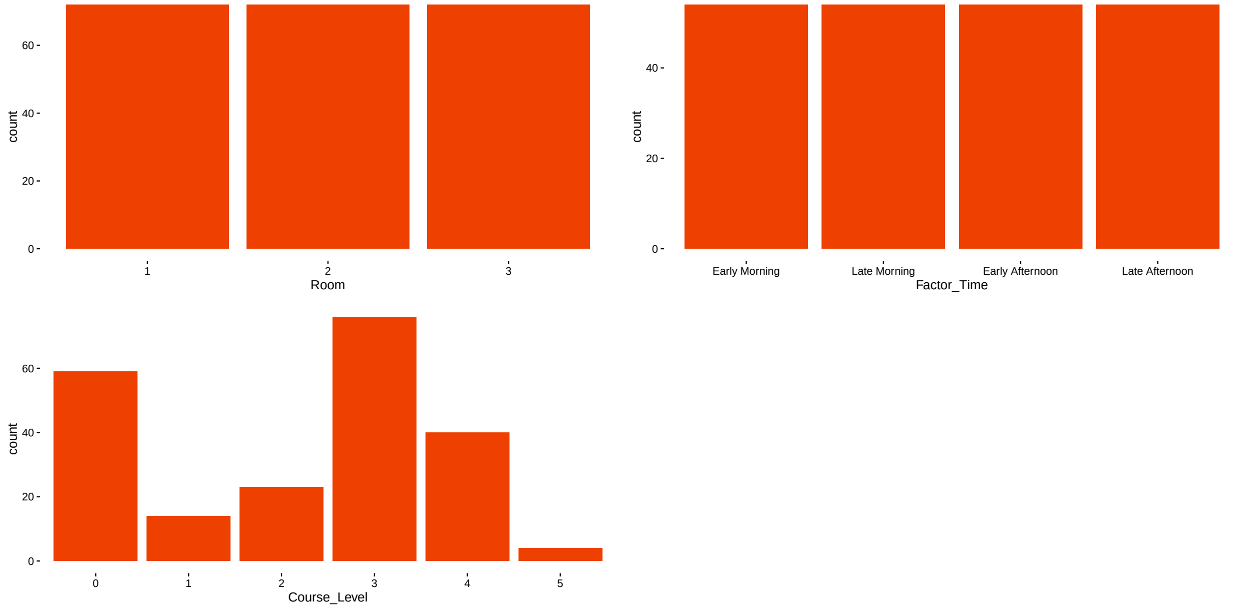
Box plots.



From the boxplots, all the trends observed in the histograms were confirmed.

For categorical variables we plot bar plot graphs.

Bar plots.



From the barplots, we could see that observations were equally distributed across all the levels of the feature Room and Factor Time. On the contrary, there were more observations for non lectures and level 3 courses. No issues were detected for those features.

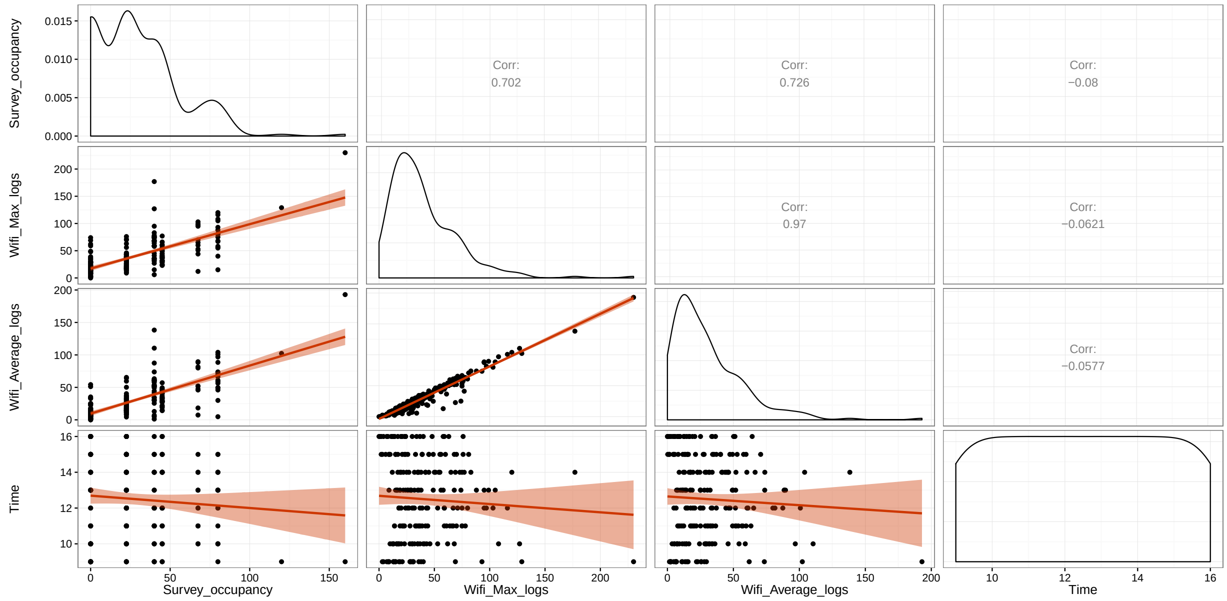
Summary.

Features	Issues	Planned Solution
Room	None	None
Time	Distibution skewed to the left	To solve during analysis
Factor Time	None	None
Course level	None	None
Tutorial	Uneven representation of the level	Discarded from the analysis
Double Module	Uneven representation of the level	Discarded from the analysis
Class went ahead	Uneven representation of the level	Discarded from the analysis
Wifi Average clients	Distibution skewed to the left & outliers	To solve during analysis
Wifi Maximum clients	Distibution skewed to the left & outliers	To solve during analysis
Survey Counted clients	Distibution skewed to the left & outliers	To solve during analysis

FEATURES AFFECTING THE TARGET FEATURE

The next step of the analysis was to see which feature really affect the target feature for deciding which features we would include into the model.
For the continuous features we explored the effects on the target features using a correlation matrix, while for the categorical features we used box plots.

Correlation matrix for continuous variables.

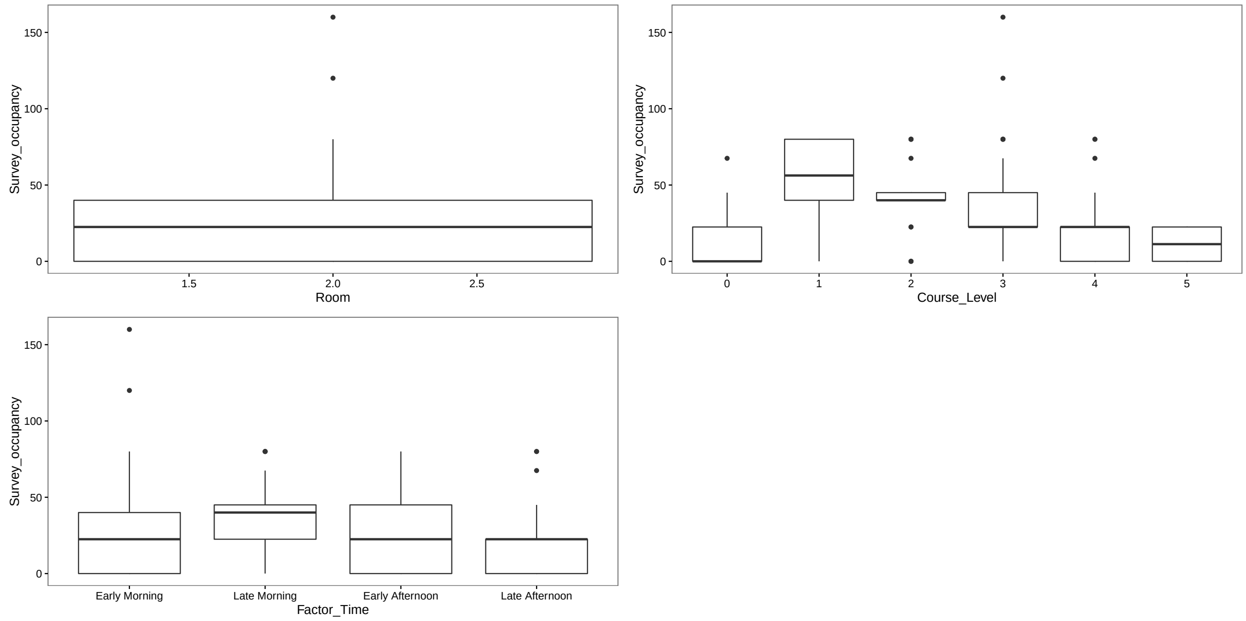


From the correlation matrix Survey counted clients seems to have a good correlation with Wifi Average counted clients and Maximum counted clients, therefore we are will try to run 2 models: one for exploring the relationship between Survey counted clients and Average counted clients and another for Survey counted clients and Maximum counted clients. However, from this graphs we can see that there is one point that is clearly two outliers. Therefore, we are going to run the analyses with and without them.

From the graphs we can see that Average counted clients and Maximum counted clients are highly correlated showing that both of them are not so different. Therefore we will not expect too much difference among the 2 models.

Time does not seems to be correlated with the target features Survey counted clients and it seems more categorical.

Box plots for categorical variables.

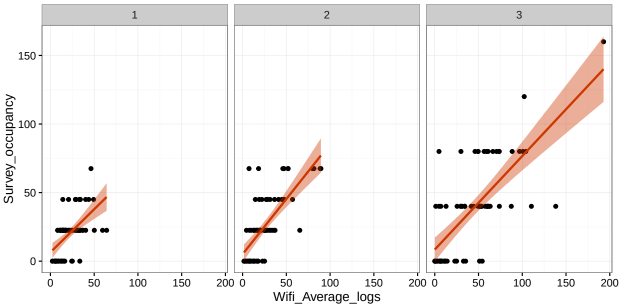


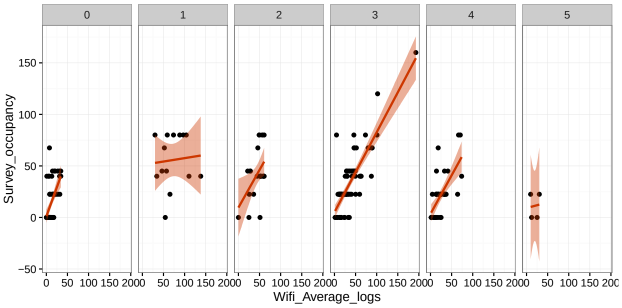
From the boxplot plotting the counted people in the different room, it can be observed that the average of the counted people in room 1 was not different from room2. Room3 had an higher number of people on average, but this can be due to the outlier. For this reasons we are going to consider its effect. The average number of counted people, instead, changed across the different level of the course and it will be worth to explore if the occupancy of the room was affected by the course level. The highest average number of counted people was in the late morning around 50, while it was around 30 for the rest of the day. Therefore it will be interesting to explore the effect of the time on the occupancy.

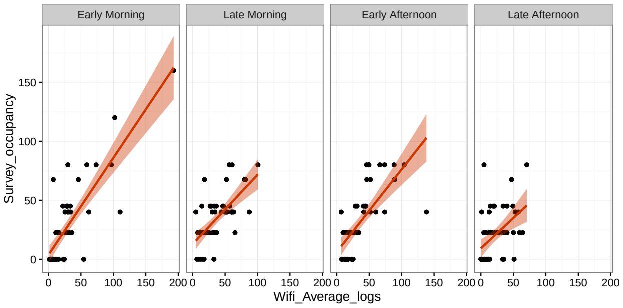
Interacting effect on the target feature

The last step before the regression was to explore the interacting effect of the features on the target features.

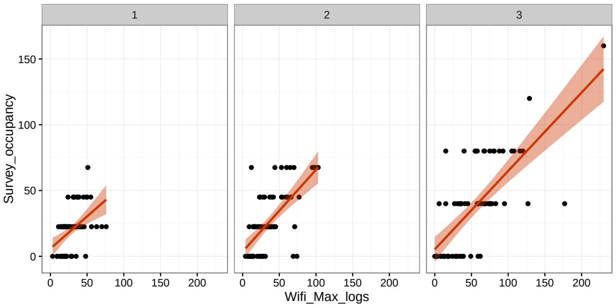
First of all we explored the interactive effect between Average counted clients with all the other features using trellis graphs.

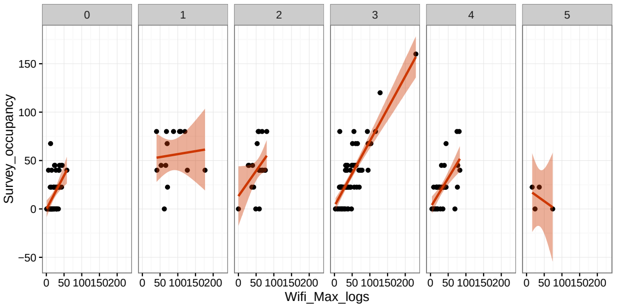


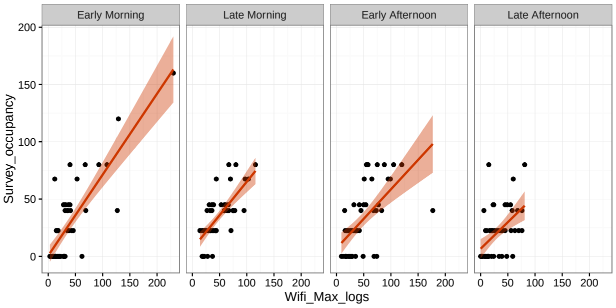




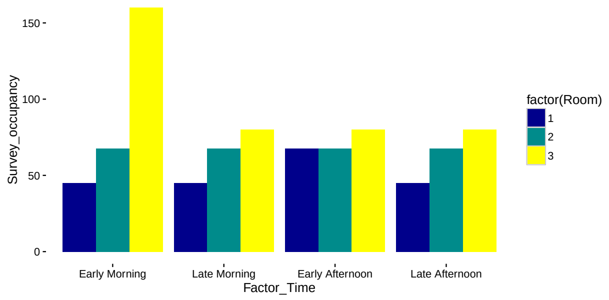
From all the graphs we could see that the positive correlation between the survey occupancy and the average wi-fi occupancy did not change either across the three room, the course levels and the time of the day, indicating that there was not interaction between these features and the average wi-fi occupancy. Similar trends were found for the same features with the maximum wi-fi occupancy (see below).

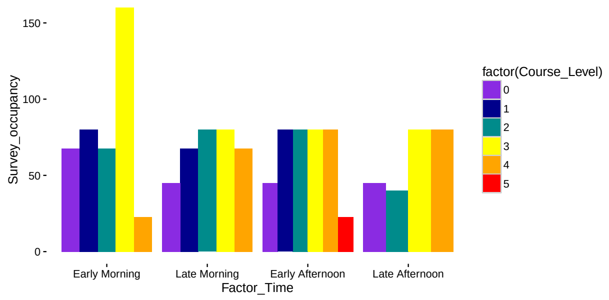


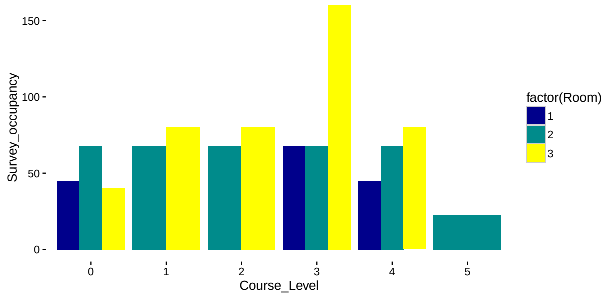




Then we explore whether categorical features were interactingly effecting the target feature using bar plots.







From the bar plots we could see that the occupancy of each room did not change across the time of the day and course level. Therefore, we are not going to consider either the interactive effect between Room and time and Course level and room. The occupancy of the different level of the courses was changing across the time of the day, suggesting a possible interaction between the course level and the time of the day.

Consequently we are going to explore the following 2 models: * Survey_occupancy ~ Wifi-Max_logs + Room + Factor_Time + Course_Level + Course_Level * Wifi-Max_logs + Factor_Time * Course_Level * Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level + Course_Level * Wifi_Average_logs + Factor_Time * Course_Level

Analysis

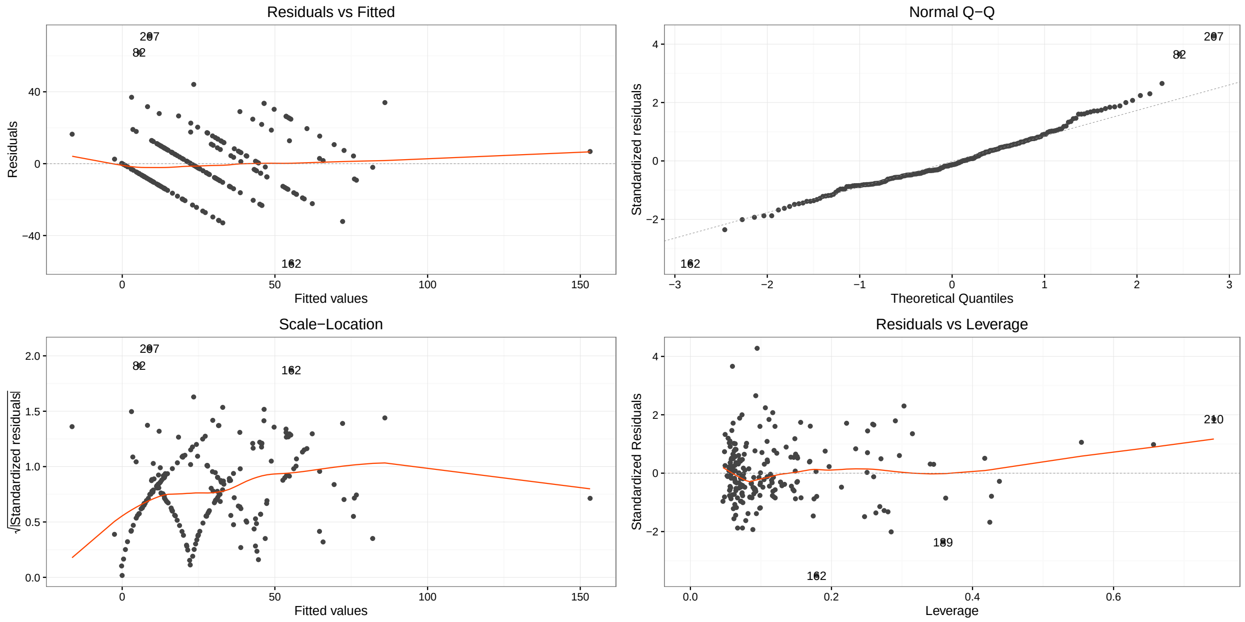
For the preliminary analysis for time constraint we run the analysis using the Validation set approach, which consists in dividing the dataset in a training and a test approach. Since the dataset was not that big, we decided to divide it in 60% for training and 40% for testing. This will give to the test dataset enough data for running the linear model. We are aware of the limitation of the Validation Set Approach and in the next analyses we are going to run the model with a 10-fold cross validation.

Models	MSE
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level+ Course_Level * Wifi-Max_logs + Factor_Time * Course_Level	21.5004107
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level + Course_Level * Wifi_Average_logs + Factor_Time * Course_Level	20.9077038

The model with the lowest MSE was the model: Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level + Course_Level * Wifi_Average_logs + Factor_Time * Course_Level. Therefore we are going to run this model on the whole dataset.

When we looked at the residuals plotted, there were few issues. As it could be seen below from the plot, showing the fitted values plotted against the residuals, the target features had a lot values closed together similarly to what expected from a categorical features and there were a potential outliers (fitted values > 140). The observations seemed normally distributed, but the variance did not seem homogeneous.

```
## [1] "ggmultiplot"
## attr(,"package")
## [1] "ggfortify"
```



For this reasons, in the future we are going to remove the outliers to see if it improve the RMSE and we are going to try to run a generalised linear model with a Poisson distribution that is usually used when dealing with counted data.